TensorFlow for Doctors*

Isha Agarwal[1][0000-0002-4116-8630], Rajkumar Kolakaluri[1][0000-0001-9696-2439], Michael Dorin[1][0000-0002-5798-0623], and Mario Chong[2][0000-0002-1231-0992]

[1] University of St. Thomas, St. Paul, MN 55105, USA
[2] Universidad del Pacífico, Lima Perú

Abstract. Machine learning has advanced substantially in the past few years, and there are many generic solutions freely available to classify text and images. The solutions are so straightforward to set up and run that having a software background is no longer necessary to perform machine learning experimentation. These systems are being adapted in many ways, and it seems only natural that those in the medical field may wish to see how machine learning might help with their research. This research examines if off-the-shelf machine learning systems are suitable for research by medical professionals who do not have software backgrounds. If all doctors who wish to experiment with machine learning could have an adequate system available, the impact on research could be substantial. This investigation applies a commonly available machine learning practice lab to medical images. As part of this investigation, we evaluated the TensorFlow for Poets (TFP) tutorial from Google Code Labs with openly available medical images provided by Kaggle Inc. While we would not recommend our test results as a basis for diagnosing medical conditions, the results were encouraging enough to suggest that using off-the-shelf systems can offer a promising opportunity to expand machine learning research into those with medical, but not software backgrounds.

Keywords: TensorFlow · Machine Learning · Image Classification.

1 Introduction

In our world there is an increasing availability of quality medical treatments for various diseases. While the pharmaceutical field continues to provide strong remedies for common ailments, it is of the utmost importance to diagnose patients early and get them the remedies they need. In first world countries it is much easier to find a doctor with adequate diagnostic technology for "one's disease." However, in developing countries where doctors are sparse and the availability of cutting-edge technology is limited, there needs to be a way to provide patients more timely diagnoses of life-threatening diseases. The purpose of this study is to research how off-the-shelf systems can allow non-technological medical professionals access to the technology they need to quickly and effectively diagnose patients. This study will examine the use of the lab "TensorFlow

* Supported by the University of St. Thomas.

for Poets" in diagnosing brain tumors, breast cancer, malaria and pneumonia. We hypothesize that it is possible for non-technical medical personnel to use off-the-shelf software and readily available hardware for providing diagnoses services or performing research. The purpose of this study is to explore this potential.

2 Relevant Existing Research

Numerous studies investigating the use of machine learning techniques in connection with diagnosis of brain tumors already exist. These can be broadly divided based on whether they use unsupervised or supervised machine learning algorithms. For example, Yassin et al. [13] use an unsupervised fuzzy clustering approach to automate brain tumor segmentation. The Brain MRI Images "activity" section of the Kaggle.com website shows the data has been viewed more than 2700 times and downloaded more than 700 times. The "kernels" and "discussion" sections show various projects based on this dataset.

Research also exists on using machine learning techniques for detection and diagnosis of breast cancer. A detailed survey on the use of machine learning algorithms on image data in the context of breast cancer can be found in Yassin et al. [13]. The BreakHist "activity" section of the Kaggle.com website shows the data has been viewed more than 1200 times and downloaded more than 200 times. The discussion section shows at least three projects based on this dataset.

Using machine learning for diagnosing malaria has been investigated in several projects. Gitonga et al. [4] present a technique for identifying the life stages and species of parasites using microscopic images of thin stained blood smears. Their paper uses an Artificial Neural Network and achieves classification accuracy of 97.76 percent in recognizing stages, and sensitivity of 93.2 percent in recognizing the species. The Kaggle.com "kernels" and "discussion" section of the pneumonia data shows various studies including the successful project "Detecting malaria cells using convolutional Neural Network" by Kushal Mahindrakar with a 92% success rate[8].

The research on using machine learning techniques for diagnosing pneumonia is not as extensive as for diagnosing other diseases such as brain tumors and breast cancer. In a recent paper, Rajpurkar et al. [11] use a 121-layer convolutional neural network (CNN) on a publicly available X-ray dataset that is labelled and consists of more than 100,000 frontal view X-rays. The pneumonia data from Kaggel.com used in our research has been used in several projects, with a main study called " Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning" [7]. The Kaggle.com "kernels" and "discussion" section of the pneumonia shows various other studies based on this data including a python based project with very high accuracy.

3 Background on TensorFlow for Poets

TensorFlow is a machine learning system which operates in different enviornments. TensorFlow provides a variety of intuitive workflows, and user-friendly application programming interfaces (API) for both beginners and experts to create machine learning models in numerous languages. TensorFlow is flexible for experimentation, and can efficiently run on various platforms ranging from mobile devices to supercomputers. This enables developers to more easily go from model building and training to deployment. This also expands the availability and utility of TensorFlow to a large audiance [6].

TensorFlow for Poets (TFP) is based on transfer learning, where instead of starting from scratch, one starts from a model that has already trained on another problem [5]. The same model is used, but the model is retrained to differentiate a small number of classes based on a new application. In contrast to deep learning from scratch, transfer learning can be done quicker.

TensorFlow For Poets makes extensive use of ImageNet. ImageNet [3] is a database created for the purpose of providing images to researchers. According to the ImageNet website, the image database is organized in a human-annotated manner and there are more than 14 million images available. The ImageNet project uses sophisticated algorithms to organize and annotate multimedia data. The availability of this large-scale image database greatly helps researchers and data scientists.

4 This Investigation

As previously mentioned, the goal of this investigation was to determine if machine learning tools could be used “off-the-shelf” for medical diagnosis or medical research. We followed the steps specified in the TensorFlow for Poets lab, just as a person with no relevant machine learning experience would do. However, instead of downloading and using the images of flowers, we downloaded and used medical images of brain tumors [2], breast cancer [12], malaria [10], and pneumonia [9] made available on the Kaggle.com website. These images were selected because of their free availability and the diversity of condition types. For retraining the network, we set the Linux shell variables exactly as indicated in the “Tensorflow for Poets lab.”

As described by Browniee’s tutorial on k–Fold Cross–Validation [1], we divided the data into groups separating a test data set from the training data. On the training data we used 10,000 training steps on brain tumor, malaria, and pneumonia images. However due to the quantity and size of the breast cancer images an increased training time was required, so for this analysis, the number of training steps was reduced to 5,000. This is because the quantity of tumor types and number of magnification levels required an increase in analysis time.

For our test data set, a minimum of 10 percent of the images were separated for validation against the trained data set. For brain tumor images, malaria, and pneumonia, the test was setup as a binary yes or no problem for both the training and the testing. TensorFlow was tasked with determining if a condition was present.

The breast cancer testing was conducted with eight different tumor types each having four zoom magnifications. This test was also binary, as we checked for one type of tumor against all other types of tumors combined. For example, for the condition of adenosis, at each zoom magnification we compared adenosis to all the remaining types to measure how successful the determination of a particular tumor type would be. For each image type and magnification, a separate training and validation was setup. The system that ran the tests is based on a Core Due 2 processor and had a GTX-970 GPU.

5 Results

A confusion matrix, commonly used in machine learning, is a table visually showing the performance of an algorithm. In this study a confusion matrix was created to demonstrate model accuracy for the brain tumor, malaria, and pneumonia analysis. Regarding Breast Cancer, we show the model accuracy for different conditions under 40X, 100X, 200X, and 400X magnifications. The overall summary of how accurate all of the non-breast cancer models were is shown in Table 2.

Confusion Matrix Brain MRI Images			
Total = 26		Predicted Class	
		Yes-Brain Tumor	No-Brain Tumor
Actual Class	Yes-Brain Tumor	12 (TP)	3 (FN)
	No-Brain Tumor	3 (FP)	8 (TN)
Model Accuracy = 77%		Misclassification Rate = 23%	

Table 1.

Beginning with the brain tumor data, some images which had been classified as positive or negative for having a brain tumor were separated before the learning process began. We then used the script provided the TFP laboratory to see how TensorFlow would classify the images. The results can be found in brain tumor section of Table 2. The confusion matrix for the brain MRI images can be found in Table 1 showing the resulting model accuracy of only 77%.

Breast cancer validation was done using the same procedures, but as there were many more files and many more types of tumors, the validation images were separated as a particular type of tumor or not that particular type of tumor. As before, we used the script provided in the TFP lab to classify the images. Table 3 shows the verification summary results for each condition. Conditions which had more data available, had better results.

Moving to the malaria validation, there were more than 1,300 non-malaria validation images and more than 1,200 malaria positive validation images. The results are in the malaria section of Table 2. The malaria confusion matrix can be found in Table 4. Overall the model accuracy was better than most, but this is likely due to the amount of data available.

	No-Brain Tumor Image	**Yes-Brain Tumor Image**
Validation Images	11	15
No Condition	8	3
Yes Condition	3	12
Model Accuracy	73%	80%
	No-Malaria Image	**Yes-Malaria Image**
Validation Images	1298	1369
No Condition	1269	252
Yes Condition	29	1117
Model Accuracy	98%	81%
	No-Pneumonia Image	**Yes-Pneumonia Image**
Validation Images	82	401
No Condition	77	13
Yes Condition	5	388
Model Accuracy	94%	97%

Table 2. Project Validation

	40X		**100X**		**200X**		**400X**	
Adenosis	13	57%	9	39%	7	32%	8	38%
Non-Adenosis	10		14		15		13	
Ductal_Carcinoma	131	76%	56	31%	104	58%	46	29%
Non-Ductal_Carcinoma	41		124		75		112	
Fibroadenoma	21	41%	14	27%	9	17%	27	57%
Non-Fibroadenoma	30		38		44		20	
Lobular_Carcinoma	5	16%	0	0%	1	3%	2	7%
Non-Lobular_Carcinoma	26		34		32		25	
Mucinous_Carcinoma	0	0%	5	11%	4	10%	2	6%
Non-Mucinous_Carcinoma	41		39		35		32	
Papillary_Carcinoma	15	52%	6	21%	5	19%	1	4%
Non-Papillary_Carcinoma	14		22		22		27	
Phyllodes_Tumor	10	45%	5	45%	4	18%	8	35%
Non-Phyllodes_Tumor	12		6		18		15	
Tubular_Adenoma	13	42%	13	43%	11	39%	6	23%
Non-Tubular_Adenoma	18		17		17		20	

Table 3. Validation Summary of BreakHist Data

Confusion Matrix Malaria Cell Images			
Total = 2667		Predicted Class	
		Yes-Malaria	No-Malaria
Actual Class	Yes-Malaria	1117 (TP)	252 (FN)
	No-Malaria	29 (FP)	1269 (TN)
Model Accuracy = 89%		Misclassification Rate = 11%	

Table 4.

With 483, pneumonia had a reasonable number of images available for final validation. Overall there were 401 classified as having Pneumonia and 82 classified as lacking pneumonia. The model accuracy is shown in Table 2. Table 5 has the confusion matrix for pneumonia.

Confusion Matrix - Pneumonia Images			
Total = 483		Predicted Class	
		Yes-Pneumonia	No-Pneumonia
Actual Class	Yes-Pneumonia	388 (TP)	13 (FN)
	No-Pneumonia	5 (FP)	77 (TN)
Model Accuracy = 96%		Misclassification Rate = 4%	

Table 5.

6 Discussion

The results do not indicate that following the procedures of our experiment would always lead to quality medical diagnoses. However, some interesting aspects still stand out. Firstly, the experiments with the most image data had a higher success rate than the categories with fewer items. Malaria and Pneumonia had many images which gave low misclassification rates. In the categories of breast cancer, ductal carcinoma had the most image samples, and also had the lowest misclassification rate.

Though the success rates were not at the diagnosis quality level, they are encouraging for future research. One step that was bypassed in this analysis was proper preparation of the data. A doctor or scientist will have the medical background to properly identify good data and remove bad data. It is important that the process has good data or good results cannot be expected. Preparing and scrubbing out the bad data is a vital process which would lead to better results.

Further research is needed to determine exactly how the system could accurately and inaccurately diagnose medical conditions. With these results, actions could be put in place to help the system work more accurately thereby decreasing inaccuracies. It is necessary to create a research team that includes medical personnel, procedural technicians, software engineers, and cultural experts. Medical personnel are needed to manually check if the system-diagnosed data was accurate. Procedural technicians (i.e. MRI and ultrasound technicians) would be integral to the staff as they could explain common errors in the images and the positioning of the patients themselves. Software engineers are necessary to run data and better the systems. Cultural experts could keep the team on task and remind the team to stay with locally accessible technology. A project manager is also recommended to help divide responsibilities and keep the team motivated for a given time line. This complete team would be best suited to providing the accuracy and care this study requires.

7 Conclusion

This paper shows the simplicity and practicality of using readily available tools to perform research for those who do not have a software or data science background. Using publicly available image data of brain tumors, breast cancer, malaria, and pneumonia, this paper investigated the reliability of applying off-the-shelf machine learning to medical diagnoses. While the accuracy of results are not currently recommended for actual medical usage, the results are encouraging to believe such technology will be useful in this way in the future. The raw data used in this study, when paired with large amounts of a disease's image data, showed a contextually low misclassification rate. Future studies on using off-the-shelf machine learning tools in the medical field are to be encouraged.

References

[1] Brownlee,Jason . A gentle introduction to k-fold cross-validation, 2018. URL https://machinelearningmastery.com/k-fold-cross-validation/. [accessed 4-July-2019].

[2] Chakrabarty, Navoneel. Brain mri images for brain tumor detection, 2019. URL https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection. [accessed 4-July-2019].

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] L. Gitonga, D. M. Memeu, K. A. Kaduki, M. A. C. Kale, and N. S. Muriuki. Determination of plasmodium parasite life stages and species in images of thin blood smears using artificial neural network. *Open Journal of Clinical Diagnostics*, 4(02):78, 2014.

[5] Google, . URL https://codelabs.developers.google.com/codelabs/tensorflow-for-poets/. [accessed 7-July-2019].

[6] Google, . URL https://www.tensorflow.org/about. [accessed 7-July-2019].

[7] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172 (5):1122–1131, 2018.

[8] Mahindrakar, Kushal. Detecting malaria cells using convolutional neural network, 2019. URL https://www.kaggle.com/kushal1996/detecting-malaria-cnn. [accessed 4-July-2019].

[9] Mooney, Paul. Chest x-ray images, 2018. URL https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. [accessed 4-July-2019].

[10] National Library of Medicine. Malaria cell images dataset, 2019. URL https://ceb.nlm.nih.gov/repositories/malaria-datasets/. [accessed 4-July-2019].

[11] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[12] Spanhol, Fabio and Eduardo Soares de Oliveira, Luiz and Petitjean, Caroline and Heutte, Laurent. A dataset for breast cancer histopathological image classification, 2015. URL https://www.kaggle.com/ankur1809/breakhist-dataset. [accessed 4-July-2019].

[13] N. I. Yassin, S. Omran, E. M. El Houby, and H. Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, 156:25–45, 2018.